Constructing Corpora of South Asian Languages

Paul Baker*, Andrew Hardie*, Tony McEnery*, and Sri B.D. Jayaram°

* Department of Linguistics, Lancaster University
° Central Institute of Indian Languages, Mysore

{j.p.baker, a.hardie, a.mcenery}@lancaster.ac.uk, jayaram@ciil.stpmy.soft.net

Abstract

The EMILLE Project (Enabling Minority Language Engineering) was established to construct a 67 million word corpus of South Asian languages. In addition, the project has had to address a number of issues related to establishing a language engineering (LE) environment for South Asian language processing, such as translating 8-bit language data into Unicode and producing a number of basic LE tools. This paper will focus on the corpus construction undertaken on the project and will outline the rationale behind data collection. In doing so a number of issues for South Asian corpus building will be highlighted.

1 Introduction

The EMILLE project[1] has three main goals: to build corpora of South Asian languages, to extend the GATE LE architecture[2] and to develop basic LE tools. The architecture, tools and corpora should be of particular importance to the development of translation systems and translation tools. These systems and tools will, in turn, be of direct use to translators dealing with languages such as Bengali, Hindi and Punjabi both in the UK and internationally (McEnery, Baker and Burnard, 2000).

This paper discusses progress made towards the first of these goals and considers to a lesser extent the third goal of the project. Readers interested in the second goal of the project are referred to Tablan et al (2002).

2 Development of the corpora

This section describes our progress in collecting and annotating the different types of corpora covered by EMILLE. EMILLE was established with the goal of developing written language corpora of at least 9,000,000 words for Bengali, Gujarati, Hindi, Punjabi, Sinhalese, Tamil and Urdu. In addition, for those languages with a UK community large enough to sustain spoken corpus collection (Bengali, Gujarati, Hindi, Punjabi and Urdu), the project aimed to produce spoken corpora of at least 500,000 words per language and 200,000 words of parallel corpus data for each language based on translations from English. At the outset we decided to produce our data in Unicode and annotate the data according to the Corpus Encoding Standard (CES) guidelines. As the project has developed, the initial goals of EMILLE have been refined. In the following subsections we describe the current state of the EMILLE corpora and outline the motives behind the various refinements that have been made to EMILLE's goals.

2.1 Monolingual written corpora

The first major challenge facing any corpus builder is the identification of suitable sources of corpus data. Design criteria for large scale written corpora are of little use if no repositories of electronic text can be found with which to economically construct the corpus. This causes problems in corpus building for the languages of South Asia as the availability of electronic texts for these languages is limited. This availability does vary by language, but even at its best it cannot compare with the availability of electronic texts in English or other major European languages.

[1] Funded by the UK EPSRC, project reference GR/N19106. The project commenced in July 2000 and is due to end in September 2003.
[2] Funded by the UK EPSRC, project references GR/K25267 and GR/M31699.

We realised that much of the data which, in principle, we would have liked to include in the corpus existed in paper form only. On EMILLE, it would have been too expensive to pay typists to produce electronic versions of the 63 million words of monolingual written corpus (MWC) data. Even if the initial typing had been affordable, checking the data for errors would have added a further cost, particularly since tools for error correction, such as spell checkers, do not exist for many of the languages studied on EMILLE (Somers, 1998, McEnery and Ostler, 2000).

Scanning in the text using an optical character recognition (OCR) program is a viable alternative to typing in printed text for languages printed in the Roman alphabet. However, OCR programs for South Asian scripts are still in their infancy (for an example of some early work see Pal and Chaudhuri, 1995) and were not considered stable and robust enough for this project to use gainfully.[3]

As part of a pilot project to EMILLE[4], we ran a workshop that examined potential sources of electronic data for Indian languages. The workshop identified the Internet as one of the most likely sources of data[5]. This prediction proved accurate, and we have gathered our MWC corpus from the web on the basis of four, largely pragmatic, criteria:

1. Data should only be gathered from sources which agreed to the public distribution of the data gathered for research purposes;

2. Text must be machine readable: we could not afford to manually input tens of millions of words of corpus data;

3. Each web-site used should be able to yield significant quantities of data: to focus our efforts we excluded small and/or infrequently updated websites from our collection effort;

4. Text should be gathered in as few encoding formats as possible: as we map all data to Unicode, we wished to limit the development of mapping software needed to achieve this.

While the first three criteria are somewhat easy to understand and have been discussed elsewhere (Baker et al, 2002) the fourth criteria merits some discussion. Ideally, we would have liked to include texts that already existed in Unicode format in our corpus. However, when we first started to collect data, we were unable to locate documents in the relevant languages in Unicode format[6]. We found that creators of such documents on the internet typically rely on five methods for publishing texts online:

- They use online images, usually in GIF or JPEG format. Such texts would need to be keyed in again, making the data of no more use to us than a paper version;
- They publish the text as a PDF file. Again, this made it almost impossible to acquire the original text in electronic format. We were sometimes able to acquire ASCII text from these documents, but were not able to access the fonts that had been used to render the South Asian scripts. Additionally, the formatting meant that words in texts would often appear in a jumbled order, especially when acquired from PDF documents that contained tables, graphics or two or more columns;
- They use a specific piece of software in conjunction with a web browser. This was most common with Urdu texts, where a separate program, such as *Urdu 98*, is often used to handle the display of right-to-left text and the complex rendering of the *nasta'liq* style of Perso-Arabic script;
- They use a single downloadable True Type (TTF) 8-bit font. While the text would still need to be converted into Unicode, this form of text was easily collected;
- They use an embedded font. For reasons of security and user-convenience, some site-developers have started to use OpenType (eot) or TrueDoc (pfr) font technology with their web pages. As with PDF documents, these fonts no longer require users to download a font and save it to his or her PC. However, gaining access to the font is still necessary for conversion to Unicode. Yet gathering such fonts is difficult as they are often protected. We found that owners of websites that used

[3] We wished to produce the corpora in the original scripts and hence avoided Romanised texts altogether.
[4] This project, Minority Language Engineering (MILLE), was funded by the UK EPSRC (Grant number GR/L96400).
[5] While we also considered publishers of books, religious texts, newspapers and magazines as a possible data source, the prevalence of old-fashioned hot-metal printing on the subcontinent made us realise early on that such sources were not likely providers of electronic data. Indeed, a number of publishers expressed an interest in helping us, but none could provide electronic versions of their texts.
[6] To date, the only site we have found that uses Unicode for Indic languages is the BBC's; see for example www.bbc.co.uk/urdu or www.bbc.co.uk/hindi.

embedded fonts were typically unwilling to give those fonts up. Consequently using data from such sites proved to be virtually impossible.

There are a number of possible reasons for the bewildering variety of formats and fonts needed to view South Asian scripts on the web. For example, many news companies who publish web pages in these scripts use in-house fonts or other unique rendering systems, possibly to protect their data from being used elsewhere, or sometimes to provide additional characters or logos that are not part of ISCII. However, the obvious explanation for the lack of Unicode data is that, to date, there have been few Unicode-compliant word-processors available. Similarly, until the advent of Windows 2000, operating systems capable of successfully rendering Unicode text in the relevant scripts were not in widespread use. Even where a producer of data had access to a Unicode word-processing/web-authoring system they would have been unwise to use it, as the readers on the web were unlikely to be using a web browser which could successfully read Unicode and render the scripts.

Given the complexities of collecting this data, we chose to collect text from South Asian language websites that offered a single downloadable 8-bit TTF font. Unlike fonts that encode English, such as Times New Roman as opposed to Courier, fonts for South Asian languages are not merely repositories of a particular style of character rendering. They represent a range of incompatible glyph encodings. In different English fonts, the hexadecimal code 42 is always used to represent the character "B". However, in various fonts which allow one to write in Devanagari script (used for Hindi among other languages), the hexadecimal code 42 could represent a number of possible characters and/or glyphs. While ISCII (Bureau of Indian Standards, 1991) has tried to impose a level of standardisation on 8-bit electronic encodings of Indian writing systems, almost all of the TTF 8-bit fonts have incompatible glyph encodings (McEnery and Ostler, 2000). ISCII is ignored by South Asian TTF font developers and is hence largely absent from the web. To complicate matters further, the various 8-bit encodings have different ways of rendering diacritics, conjunct forms and half-form characters. For example, the Hindi font used for the online newspaper *Ranchi Express* tends only to encode half-forms of Devanagari, and a full character is created by combining two of these forms together. For example, to produce he (Unicode character U+092A) in this font, two keystrokes would need to be entered (*h* + *e*).However, other fonts use a single keystroke to produce he.

We were mindful that for every additional source of data using a new encoding that we wished to include in our corpus, an additional conversion code page would have to be written in order to convert that corpus data to the Unicode standard. This issue, combined with the scarcity of electronic texts, meant that we didn't use as many sources of data as we would have initially liked. Thus we had to focus almost exclusively on newspaper material[7]. However, as noted in the following paragraph, as a consequence of the collaboration between Lancaster University and the Central Institute of Indian Languages (CIIL), the eventual corpus will now contain a wider range of genres.

Web data gathered on the basis of these four criteria would have allowed us to fulfil our original MWC project goals. However, the MWC collection goals of the project have altered significantly. Thanks to a series of grants from the UK EPSRC[8] the EMILLE project has been able to establish a dialogue with a number of centres of corpus building and language engineering research in South Asia. As a consequence, the EMILLE team has joined with the CIIL in Mysore, India, to produce a wider range of monolingual written corpora than originally envisaged on the EMILLE project. One effect of this change is that the uniform word counts of the monolingual written corpora will be lost.[9] Each language will now be provided with varying amounts of data, though no language will be furnished with less than two million words. However, there is a further important effect of this collaboration: the corpus now covers a much wider range of languages (14 rather than 7) and a wider range of genres. By a process of serendipity, the corpus data provided by CIIL covers a number of genres, but not newspaper material.[10] As the material gathered at Lancaster focuses almost exclusively on newspapers, the CIIL

[7] One important exception to this is the incorporation of the Sikh holy text, the Adi Granth or Guru Granth Sahib, into the Punjabi corpus.
[8] Grants GR/M70735, GR/N28542 and GR/R42429/01.
[9] This change was also necessitated by the varying availability of suitable newspaper websites for the different languages. For Hindi and Tamil, for example, plenty of data is available to be gathered; for Punjabi and Bengali, somewhat less; for Urdu, almost none.
[10] The data provided by CIIL to the project covers a number of genres, including Ayurvedic medicine, novels and scientific writing.

and Lancaster data is complementary. Table 1 shows the size of the EMILLE/CIIL monolingual written corpora.

Language	***Written Corpus Size in words (millions)***
Assamese	2,620,000
Bengali	5,520,000
Gujarati	12,150,000
Hindi	12,390,000
Kannada	2,240,000
Kashmiri	2,270,000
Malayalam	2,350,000
Marathi	2,210,000
Oriya	2,730,000
Punjabi	15,600,000
Sinhalese	6,860,000
Tamil	19,980,000
Telegu	3,970,000
Urdu	1,640,000
Total	90,172,000

Table 1: Word counts for the written part of the EMILLE/CIIL Corpus

2.1.1 Encoding of the monolingual written corpora

The decision had been made early on to use CES encoding for the EMILLE corpora. However, for CES documents there are a variety of different levels of conformance. Level 2 conformance requires the use of SGML entities to replace special characters such as emdashes, currency signs, and so on; however, since the corpus was to be encoded as 16-bit rather than 7-bit text this was not a matter of importance, as there is provision in the Unicode standard for all these "special" characters. Level 3 conformance involves adding tags for abbreviations, numbers, names and foreign words and phrases. Given the size of the corpus, this could only be practicable if accomplished automatically; however, given the large range of writing systems we were dealing with, a suitable algorithm would have been too time-consuming to implement.

The corpus texts therefore extend only to level 1 CES conformance. For this it is necessary for documents to validate against the cesDoc DTD. The level 1 recommendation that italics and similar textual information that might conceivably indicate a linguistically relevant element be retained has not been implemented, again due to the unwieldiness of attempting to do so for the large number of scripts involved.

In practice, this means that for monolingual written texts the main textual elements are <p>, <s>, and <head>. These elements could be deduced automatically from the HTML code of the original web pages from which the data was gathered. A full CES header has been used, however; see the Appendix to this paper for an example.

2.1.2 Mapping the corpus texts to Unicode

The task of mapping the data in the monolingual written corpus is, as has been indicated above, a fairly difficult one. Whilst it is fairly simple to write a program that will map every character in a given font to one or more given Unicode characters, this basic algorithm will not handle the more problematic fonts. The formats we had to deal with fell into three broad groups.

- Texts in Urdu or western Punjabi required one-to-one or one-to-many character mapping. This was due to the nature of the alphabet[11] in which in they were written, which does not contain conjunct consonants as the Indian alphabets do.
- Texts in ISCII required one-to-one character mapping. These texts, primarily those from the data provided by CIIL, could be mapped very simply because the Unicode standard for Indian alphabets is actually based on an early version of the ISCII layout.
- Texts in specially-designed TTF fonts as discussed above required the most complex mapping. They typically contain four types of characters. The first type need to be mapped to a string of one or more Unicode characters as with ISCII and the Perso-Arabic script. The second type have two or more potential mappings, conditional on the surrounding characters. Some of these conditional mappings could be handled by generalised rules; others operated according to character-specific rules. The third type of characters required the insertion of one or more characters into the text stream prior to the point at which the character occurred[12]. The fourth type, conversely, required characters to be inserted into the text stream *after* the current point (in effect, into a Unicode stream which does not yet exist)[13]. In neither of the latter two types was it simply a case of going "one character forwards" or "one character back"; the insertion point is context-sensitive.

The third type of text in particular could not be dealt with using simple mapping tables – each font required a unique conversion algorithm. The task of developing and coding these algorithms was split between a partner in India and the University of Lancaster. The "Unicodify" software suite developed at Lancaster is currently capable of mapping three fonts for three separate languages, and has been successfully used to encode the monolingual data published in the beta release of the corpus (see section 4.0 below).

Unicodify is also capable of re-interpreting the HTML elements created by the "save as Web Page" function of Microsoft Office programs and mapping them to the CES elements of <p>, <head> and <s>, and of generating an appropriate header.

At the time of writing, the collection phase for the EMILLE/CIIL MWC data is nearly complete. Only around 13 million words of data remain to be collected (as shown in Table 1 above). Good progress is also being made on mapping this data to Unicode. Consequently, the focus of the project is now falling increasingly on parallel and spoken data.

2.2 Parallel corpora

The problems we encountered in collecting MWC data were also encountered when we started to collect parallel data. However, the relatively modest size of the parallel corpus we wished to collect (200,000 words in six languages) meant that we were able to contemplate paying typists to produce electronic versions of printed parallel texts. We eventually decided to do this as we had an excellent source of parallel texts which covered all of the languages we wished to look at: UK government advice leaflets. This was a good source of data for us, as we wished to collect data relevant to the translation of South Asian languages in UK in a genre that was term rich.

The leaflets we were able to gather were mostly in PDF or print-only format. Typing these texts became a necessity when the UK government gave us permission to use the texts, but the company that produced the electronic versions of the texts refused to give us the electronic originals. We found it

[11] We have come to refer to this alphabet as "Indo-Perso-Arabic", although it is more widely known simply as the "Urdu alphabet", or in the case of the various forms of western Punjabi that use it, "Shahmukhi". This name is designed to capture the fact that the Perso-Arabic script as used for Indo-Aryan languages has certain shared features not found in Arabic, Persian, etc. – for instance, characters for retroflex consonants, or the use of the *nasta'liq* style of calligraphy.

[12] This is primarily the case for those Indian alphabets which allow conjunct consonants whose first component is the letter "ra". When this letter is the first half of a conjunct, it takes the form of a diacritic which appears *after* the second half of the conjunct. In Unicode, the text stream contains the logical order of the characters, but in the TTF fonts, the graphical order is almost always the order that is held in the computer's memory.

[13] This is primarily the case for certain vowel diacritics which indicate vowels that follow the consonant but which appear before the consonant. Again, Unicode follows the logical order, whereas TTF fonts almost always follow the graphical order of the glyphs.

economic to pay typists to produce Unicode versions of the texts using Global Writer, a Unicode word-processor.[14]

The research value of the British government data is very high in our view. The UK government produces a large number of documents in a wide range of languages. All are focused in areas which are term-rich, e.g. personal/public health, social security and housing. To build the parallel corpus we collected 72 documents from the Departments of Health, Social Services, Education and Skills, and Transport, Local Government and the Regions.[15]

Other than the need to type the data from paper copies, the parallel corpus also presented one other significant challenge: while most of the data is translated into all of the languages we need, there are a few instances of a document not being available in one of the languages. Our solution is to employ translators to produce versions of the documents in the appropriate language. While far from ideal, this is not unprecedented as the English Norwegian Parallel Corpus project also commissioned translations (see Oksefjell, 1999). All such texts are identified as non-official translations in their header.

The parallel corpus is now complete, and we are beginning the process of sentence aligning the texts using the algorithm of Piao (2000).

2.2.1 Annotation of the parallel corpora

The annotation of the parallel texts was essentially the same as that applied to the monolingual texts. The principal difference was that the parallel texts had to be keyboarded manually by native speakers of the relevant languages, and thus rather than the automated insertion of SGML elements which characterises the monolingual written corpora, it was necessary to formulate guidelines that would allow transcribers with no knowledge of SGML to accurately mark up the text. In the event, the ultimate content of the transcription guidelines was dictated by certain problems relating to the recruitment of transcribers.

Our initial strategy was to recruit typists from among the student body of Lancaster University, which includes native speakers of all the languages in question (Hindi-Urdu, Bengali, Punjabi, and Gujarati). However, we were only able to find two or three reliable transcribers in this way. The majority of students recruited were unable to commit themselves to working on the project for the necessary extended period. This problem was particularly acute because we had to recruit outside the Department of Linguistics; while there were potential academic benefits for Linguistics students which would make the task more worthwhile, this was not the case for students of other subjects.

Therefore, we have formed an arrangement with a data-processing firm in India, Winfocus PVT, who have been able to take on the transcription of a significant proportion of the parallel corpus (as well as nearly all the spoken data – see also below). However, because Winfocus deployed a wide range of staff members on the project, it was necessary that the transcription guidelines be as simple as possible, as typing would otherwise be slowed to an unacceptable speed.

The decision had already been made for the monolingual corpora to adhere to only the most basic level of CES-compliance (see above). This meant straightaway that the guidelines could be fairly simple. The SGML elements included in all versions of the guidelines were <s>, <head>, <p>[16], and <foreign> – these were also used in the monolingual corpus. Unique to the parallel corpus was the tag <corr> for a transcriber's correction of a typographical error in the original printed text (the SGML of course retains the original "incorrect" version).

The instructions for the <gap> element (replacing pictures and other non-transcribable graphical elements) have likewise always been part of the guidelines. It would seem that these instructions have

[14] When the project began, Global Writer was one of the few word-processors able to handle the rendering of Indic languages in Unicode. Since then, Microsoft have made Word 2000 Unicode-compliant. However, unless running on a Windows 2000 machine the Unicode compliance of Word 2000 is not apparent.
[15] We also collected a smaller number of texts from the Home Office, the Scottish Parliament, the Office of Fair Trading, and various local government bodies (e.g. Manchester City Council).
[16] Lists of bulleted-pointed items – which are very common in the UK government information leaflets – are encoded as separate <s> elements within a single <p>, as are table cells.

been ignored by transcribers, however, who have just passed silently over the illustrations[17]. We intend to harmonise the use of the <gap> element in the data at the end of the project.

In the initial version of the guidelines, instructions were included for an SGML encoding of footnotes using <ptr> elements to anchor the notes and <note> elements at the end of the file containing the footnote text. However, in practice transcribers seem to have ignored these guidelines as well, encoding footnotes using normal <p> and <s> elements wherever they physically occurred in the text. Therefore <ptr> and <note> were excised from later versions of the guidelines, as were the instructions for the <bibl> reference for bibliographic references – which in the event was never needed anyway.

Initially, we also asked the transcribers to add details of title, publication date/place, etc. from the printed text to the header of the file as they typed it. However, due to regular inconsistencies and errors in the headers thus created, we later moved over to a system in which the header was added subsequently by a member of the EMILLE research team using information from a central database. Doing this drastically reduced the bulk of the transcribers' guidelines, greatly facilitating the training of new transcribers.

Similarly, it is a design feature of the EMILLE corpora that the filename of each text embodies the hierarchical structure of the corpus as a whole. For the parallel corpus, this means that each filename includes information on the language, medium and category of the text, as well as a descriptive name which is that text's unique identifier. For example, the Gujarati version of the text "The Health of the Nation and You", an NHS information booklet published by the Department of Health, has the filename **guj-w-health-nation.txt**[18]. The guidelines initially gave step-by-step instructions for composing these filenames, but transcribers uniformly got it wrong, so this task too was excised from later versions of the guidelines.

2.3 Spoken corpora

For the collection of spoken data we have pursued two strategies. Firstly we explored the possibility of following the BNC (British National Corpus) model of spoken corpus collection (see Crowdy, 1995). We piloted this approach by inviting members of South Asian minority communities in the UK to record their everyday conversations. In spite of the generous assistance of radio stations broadcasting to the South Asian community in the UK, notably BBC Radio Lancashire and the BBC Asian Network, the uptake on our offer was dismal. One local religious group taped some meetings conducted in Gujarati for us, and a small number of the people involved in transcription work on the project agreed to record conversations with their family and friends. The feedback from this trial was decisive – members of the South Asian minority communities in Britain were uneasy with having their everyday conversations included in a corpus, even when the data was fully anonymised. The trial ended with only 50,000 words of spoken Bengali and 40,000 words of Hindi collected in this way.

Consequently we pursued our second strategy and decided to focus on Asian radio programmes broadcast in the UK on the BBC Asian Network as our main source of spoken data.[19] The BBC Asian Network readily agreed to allow us to record their programmes and use them in our corpus. The five languages of the EMILLE spoken corpora (Bengali, Gujarati, Hindi-Urdu, and Punjabi) are all covered by programmes on the BBC Asian Network. At least four and a half hours in each language (and more in the case of Hindi-Urdu) are broadcast weekly. The programmes play Indian music (the lyrics of which have not been transcribed) as well as featuring news, reviews, interviews and phone-ins. As such the data allows a range of speakers to be represented in the corpus, and some minimal encoding of demographic features for speakers is often possible as at least the sex of the speaker on the programmes is apparent.

The recordings of the radio programmes are currently being digitised and edited, to remove songs and other such material. The recordings will be made available in conjunction with the transcriptions. However, the transcriptions and recordings will not be time aligned. An obvious future enhancement of

[17] It should however be noted that one transcriber went the opposite way, being extremely "trigger happy" with his use of the <gap> element and filling the text with non-requisite information which had to stripped later.
[18] Spoken texts are similarly titled; for instance the file **guj-s-cg-asiannet-02-11-23.txt** is a context-governed spoken text containing a transcription of the BBC Asian Network Gujarati programme transmitted on the 23rd November 2002.
[19] Programmes broadcast in Bengali and Urdu on BBC Radio Lancashire make up the remainder of the spoken corpus.

this corpus data would be to work on techniques, already well established for English, to time align the transcriptions.

The recording and transcription of the broadcasts is complete and the size of the EMILLE spoken corpora are shown in Table 2 below.

Language	***Written Corpus Size in words (millions)***
Bengali	442,000
Gujarati	564,000
Hindi	588,000
Punjabi	521,000
Urdu	512,000
Total	2,627,000

Table 2: Word counts for the spoken part of the EMILLE/CIIL Corpus

2.3.1 Annotation of the spoken corpus

The transcription of the spoken texts was undertaken by the same group of typists who worked on the parallel corpus. In a similar way, the annotation guidelines, which began by embracing the full range of possible CES encoding elements, were of necessity simplified as the project progressed.

As with the parallel transcriptions, the requirement to generate an appropriate header and filename were dropped early on. Information required for the header – on participants sex/age/profession, their relationships, and the setting of the conversation – was noted in plain English at the top of the file by the transcribers, and then converted to a proper SGML header later on.

Within the text itself, the elements that were implemented were <u> for utterance – which also records the speaker and a unique ID for each utterance – and also <foreign>, <unclear>, <omit>, <vocal>, and <event>, which relate to either the content of the speech or other noises on the tape. The codes for noting overlapping speech, the <pause> code, and the large set of <shift> codes to indicate changes in stress, seem to have been almost entirely neglected by the transcribers, so although they were not removed from the guidelines their importance was de-emphasised in later versions to avoid distracting attention from the greater importance of the <u>, <unclear> and <omit> elements.

3 Analytic annotation of the corpora

We aimed from the outset to explore morphosyntactic annotation of the Urdu corpus. For a description of the work undertaken on this aspect of the project, see Hardie (2003).

The corpus annotation research of EMILLE has recently expanded to cover another form of annotation – the annotation of demonstratives – in Hindi. The work on Hindi is at an early stage, with an annotation scheme originally designed to annotate demonstratives in English (Botley and McEnery, 2001) being used to annotate Hindi. The annotation is currently underway and the goal is to annotate the demonstratives in 100,000 words of Hindi news material by the end of the project.

4 Accessing the corpus

A beta release of the EMILLE/CIIL corpus will be available, free of charge, for users from April 2003. The beta release of the corpus will contain a sample of MWC, parallel and spoken data for the core EMILLE languages. In order to register for access to the beta release, users should contact Andrew Hardie.

5 Conclusion

The EMILLE project has adapted and changed over the course of the past two years. With regard to the EMILLE corpora, this has in large part been due to the project team engaging in a dialogue with the growing community of researchers working on South Asian languages. As a result of this dialogue the EMILLE team has made some major changes to the original design of the EMILLE corpora. However, as with all large-scale corpus-building projects, other changes have occurred on the project which have

been responses to unexpected factors, such as the reluctance of members of the minority communities to engage in the recording of everyday spontaneous speech, and the lack of compatible 8-bit font encoding standards used by the different producers of electronic texts in the relevant languages. Devising methodologies to convert the numerous disparate 8-bit based texts to Unicode has been one of the most complex and time-consuming tasks of the project.

The area of South Asian corpus building is growing. As well as work in the UK and India, a new centre for South Asian language resources has been established in the US[20]. As the centres cooperate and integrate their research, there is little doubt that further work on the construction and annotation of South Asian corpora will grow. As this work grows, we believe that corpus builders should not loose sight of two important truths. Firstly, that collaboration is better than competition – the corpus produced by Lancaster/CIIL will be larger and better because we have accepted this. The construction of large scale language resources needs the acceptance of this truth if it is to be effective. Secondly, that while many South Asian languages are entering the growing family of languages for which corpus data is available, there are still languages spoken in South Asia and the world for which corpus data is not available. While we must celebrate the creation of corpora of South Asian languages, we should also think of the work yet to be done in creating corpora for those languages not yet corpus enabled.

References

Baker, JP, Burnard, L, McEnery, AM and Wilson, A 1998 Techniques for the Evaluation of Language Corpora: a report from the front. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada.

Baker, JP, Hardie, A, McEnery, AM, Cunningham, H, and Gaizauskas, R 2002 EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In: González Rodíguez, M and Paz Suarez Araujo, C (eds.) *Proceedings of 3rd Language Resources and Evaluation Conference(LREC)*. Las Palmas de Gran Canaria.

Botley, S.P. and McEnery, A 2001 Demonstratives in English: a Corpus-Based Study. *Journal of English Linguistics*, 29: 7-33.

Bureau of Indian Standards (1991) *Indian Standard Code for Information Interchange*, IS13194.

Crowdy, S 1995 The BNC spoken corpus. In: Leech, G, Myers, G and Thomas, J (eds.), *Spoken English on computer: transcription, mark-up and application*. Longman: London.

Hardie, A 2003 Developing a model for automated part-of-speech tagging in Urdu. Paper presented to CL2003 conference.

McEnery, A, Baker, JP and Burnard, L 2000 Corpus Resources and Minority Language Engineering. In M. Gavrilidou, G. Carayannis, S. Markantontou, S. Piperidis and G. Stainhauoer (eds) *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*: Athens.

McEnery, AM and Ostler, N 2000 A New Agenda for Corpus Linguistics – Working With All of the World's Languages. In *Literary and Linguistic Computing*, 15: 401-418.

Oksefjell, S 1999 A Description of the English-Norwegian Parallel Corpus: Compilation and Further Developments. In *International Journal of Corpus Linguistics*, 4:197-219.

Pal, U and Chaudhuri, BB (1995) Computer recognition of printed Bengali script. In *International Journal of System Science*, 26: 2107-2123.

Piao, SS 2000 *Sentence and Word Alignment Between Chinese and English*. Ph.D. thesis, Department. of Linguistics and Modern English Language, Lancaster University, UK.

Somers, H 1998 Language Resources and Minority Languages. *Language Today* 5.

Tablan, V, Ursu, C, Bontcheva, K, Cunningham, H, Maynard, D, Hamza, O and Leisher, M 2002 A Unicode-based Environment for Creation and Use of Language Resources. In: González Rodíguez, M and Paz Suarez Araujo, C (eds.) *Proceedings of 3rd Language Resources and Evaluation Conference(LREC)*. Las Palmas de Gran Canaria.

[20] See http://ccat.sas.upenn.edu/~haroldfs/pedagog/salarc/overallplan.html

Appendix

Example of the CES header for a monolingual corpus text. Note that throughout the corpus the date format *yy-mm-dd* is employed.

```
<cesDoc id="guj-w-samachar-news-01-05-23" lang="guj">
<cesHeader type="text">
<fileDesc>
<titleStmt>
<h.title>guj-w-samachar-news-01-05-23.txt</h.title>
<respStmt>
<respType>Electronic file created by</respType>
<respName>Department of Linguistics, Lancaster University</respName>
<respType>text collected by</respType>
<respName>Andrew Hardie</respName>
<respType>transferred into Unicode by</respType>
<respName>"Unicodify" software by Andrew Hardie</respName>
</respStmt>
</titleStmt>
<publicationStmt>
<distributor>UCREL</distributor>
<pubAddress>Department of Linguistics, Lancaster University, Lancaster, LA1 4YT, UK</pubAddress>
<availability region="WORLD"></availability>
<pubDate>02-12-18</pubDate>
</publicationStmt>
<sourceDesc>
<biblStruct>
<monogr>
<h.title>"Gujarat Samachar" internet news (www.gujaratsamachar.com), news stories collected on 01-05-
23</h.title>
<h.author>Gujarat Samachar</h.author>
<imprint>
<pubPlace>Gujarat, India</pubPlace>
<publisher>Gujarat Samachar</publisher>
<pubDate>01-05-23</pubDate>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc>Text collected for use in the EMILLE project.</projectDesc>
<samplingDesc>Simple written text only has been transcribed. Diagrams, pictures and tables have been omitted
and their place marked with a gap element.
</samplingDesc>
<editorialDecl>
<conformance level="1"></conformance>
</editorialDecl>
</encodingDesc>
<profileDesc>
<creation>
<date>02-12-18</date>
</creation>
<langUsage>Gujarati</langUsage>
<wsdUsage>
<writingSystem id="ISO/IEC 10646">Universal Multiple-Octet Coded Character Set (UCS).</writingSystem>
</wsdUsage>
<textClass>
<channel mode="w">print</channel>
<constitution type="composite"></constitution>
<domain type="public"></domain>
<factuality type="fact"></factuality>
</textClass>
<translations></translations>
</profileDesc>
<revisionDesc></revisionDesc>
</cesHeader>
```